

「モデル選択」の前後

石黒 真木夫

1. 「モデル選択」の前

表1は、統計数理研究所が1983年の8月末に開催した公開講座の出席者数のデータである。筆者も講師の1人を務めたこの講座は8月22日から31日まで、日曜日をのぞいて9日間にわたって開かれた。表の数字は最初の5日間分である。5日分しかないのは、このデータを解析してみようと思ったのが6日目のことで、5日目までの分か集計されていなかったからである。

出席者数が減少する傾向にあるか否かに興味があった。そしてまた、減少しつつあるというのなら9日目には何人くらいにまで落ちるものか予想しておきたかった。この講座は筆者がその著者の1人になっている本を使ってのはじめての講義だった。従来の統計学書とはかなり趣きを異にする書き方をした本だったから、聴衆が最後までついてきてくれるか心配だったのである。

こう考えた。80人近い人が集まっているからには、さまざまな事情で毎日は出席できない人も含まれているだろう。これら未知の要因による出席者数の増減は講座の出来・不出来とは関係ない。気になるのはこの偶然変動をとり除いた“本質的な出席者数”の動きである。つまり“本質的な出席者数が減少傾向にあるか否か”が問題である。

統計学ではこのような場合、“本質的な出席者数

表1 5日目までの出席者数

日付	8/22	23	24	25	26
人数	78	72	76	74	73

が減らないモデル”と、“本質的な出席者数が減るモデル”の2つを用意して、“どちらのモデルが良いか”という形に定式化するのが普通である。すなわち「モデル選択」の問題としてとり扱う。

ここでいうモデルは“統計的モデル”であり、データを確率変数の実現値として説明するモデルである。

まず、本質的な出席者数が減らないモデルとして、少々乱暴だが、

$$(1) \quad n_i = \mu + \varepsilon_i$$

を採用しよう。ここで n_i は i 日目の出席者数、 μ は i によらない(未知の)定数である。 ε_i には、 n_i の偶然変動分として、平均0の正規分布を仮定する。 n_i は整数値をとるはずだから ε_i に正規分布を仮定するのは冒険だが、眼をつぶる。 $i \neq j$ なら ε_i と ε_j は独立とする。

本質的な出席者数が減るモデルとしては、

$$(2) \quad n_i = \mu_0 + \mu_1 \alpha^i + \varepsilon_i$$

を採用しよう。 $0 < \alpha < 1$ なら、 n_i の平均は初日 $i=1$ の $\mu_0 + \mu_1 \alpha$ から徐々に減少して $i \rightarrow \infty$ で μ_0 になる。 ε_i はモデル(1)の場合と同じに定義する。

現実の出席者数の増減の裏には非常に複雑なメカニズムが働いているにちがいない。その複雑さを一切捨象した(1)式や(2)式をいくらよく見て

も、生身の受講者の熱心にうなずく顔やあくびをこらえる口もとは見えてこない。しかし(1)式や(2)式には“とりえ”がある。単純なのである。

こう定式化してはじめて“出席者が減少する傾向にあるか否か”という疑問が、“モデル(1)とモデル(2)ではどちらのモデルが良いモデルか”という統計的な問題になる。

これはほんの一例にすぎないが、われわれがデータ(あるいはデータをとって調べてみたいと思う現象)を前にして感ずる素朴な疑問を統計的なモデル選択の問題に翻訳できることが多い。

しかし、日本語から英語に翻訳するやり方が一意には決まらないのと同じ意味で、日常語の疑問文の統計学語訳は一意に決まらない。たとえば、“出席者数が減る”ことの表現はモデル(2)の他にもあり得る。実際、筆者が最初に考えたモデルは(2)ではなく、

$$(2)' \quad n_i = \mu_1 \alpha^i + \epsilon_i$$

だった。

モデル(2)'を日常語に逆翻訳してみると、 $0 < \alpha < 1$ の場合、“出席者は次第に減って、ついにはだれもいなくなる”ということになるのか。あまりにも非現実的な(というよりは、あまりにも悲しい)モデルなので(2)の形を採用することにしたのである。

一般に、統計的な「モデル選択」によって問題を解こうとするときに、最もむずかしく、そしてまた最もおもしろいのは、この日常語から統計学語への翻訳である。翻訳の常で誤訳の可能性がつきまとう。正確な翻訳には、日常語(つまりは問題意識)と統計学語(つまり統計的な情報処理のテクニック)に関する知識とセンスが不可欠である。多くの場合、対象とする分野の専門家と統計学者の緊密な協力があってはじめてこれが可能になる。

2. 統計的「モデル選択」

モデル選択の手段として用いられる方法に“検

定法”と“規準法”がある。2つの物を“てんびん”の左右に乗せてその傾きを見る方法と、目方を比べる方法の2通りがあるのに似ている。てんびんに相当するのは検定法のほうである。

ここでは統計数理研究所の赤池氏によって導入された、典型的な規準法である、情報量規準AICを紹介する。(参考文献としては赤池氏自身の論文[1, 2, 3]と坂元、石黒、北川共著の[4]をあげておく。最初に表1のデータの説明のところであられた公開講座で使った“本”というのは実は[4]である)

詳しいこと、特にその導き方は参考文献にゆずるが、AICの簡単な説明をしておこう。

まず、AICはほとんどすべての統計学の教科書にのっている“最尤法”と深い関係があり、

$$(3) \quad AIC = (-2) \ln(\text{最大尤度}) + 2(\text{パラメータ数})$$

で定義される。ここで最大尤度というのは“尤度関数”の最大値であり、仮定したモデルに対してユニークに定まる。“パラメータ数”はそのモデルに含まれる自由なパラメータの数である。とにかく統計的モデルのパラメータを最尤法で推定すれば、AICの値が確定するのである。

いくつかモデルがあるとき、AICの値が小さいほど(いま問題にしているデータに関しては)“より良い”モデルと考えられる。ボルツマンの意味のエントロピーを基本的な評価規準にとると“モデルの良さ”を表現する客観的な量(=期待平均対数尤度)が定義される。AICの(-1/2)倍が期待平均対数尤度の不偏推定量となるのである。期待平均対数尤度が大きいほどより良いモデルである。上でAICが小さいほどより良いモデルと考えられるとぼかしたのは、AICが本当の規準の推定量だからである。

AICを用いて、モデル(1)とモデル(2)ではどちらが良いか調べてみよう。

モデル(1)の尤度関数は、 $\{\epsilon_i\}$ が互いに独立な正規分布にしたがうとの仮定から、

$$(4) L(\mu, \sigma^2) = \prod_{i=1}^5 \frac{1}{\sqrt{2\pi} \sigma} e^{-(n_i - \mu)^2 / 2\sigma^2}$$

となる。 σ^2 は $\{\epsilon_i\}$ の分散であり、これと μ の2個がモデル(1)の自由なパラメータということになる。(4)式の $\{n_i\}$ に表1のデータを与えて最尤法を適用することによってAICの第1項が決まり、“パラメータ数”が2であることから、結局モデル(1)のAICは、

$$(5) AIC_1 = 25.9$$

と計算される。

モデル(2)の自由なパラメータは μ_0, μ_1, σ^2 および α の4個である。これらのパラメータに関する尤度関数は、

$$(6) L(\mu_0, \mu_1, \sigma^2, \alpha) = \prod_{i=1}^5 \frac{1}{\sqrt{2\pi} \sigma} e^{-(n_i - \mu_0 - \mu_1 \alpha^i)^2 / 2\sigma^2}$$

となり、これの最大値と“パラメータ数”=4から得られるモデル(2)のAICの値は、

$$(7) AIC_2 = 25.0$$

である。

AICの値は小さいほうが良い。25.0は25.9より小さい。AIC₂がAIC₁より小さい。すなわちモデル(1)よりモデル(2)のほうがより良いモデルと考えられる。

3. 「モデル選択」の後

表1のデータに関してはモデル(1)よりモデル(2)のほうが良いと考えられるということになった。出席者数は減少傾向にあると結論してよいか。われわれの講座は失敗だったのだろうか。モデル(2)のAICは25.0だった。この値を計算するにあたって最大尤度を求めた。最大尤度を与えたパラメータの値、“最尤推定値”を見ると、

$$(8) \hat{\mu}_0 = 73.75, \hat{\mu}_1 = 4248.87, \hat{\alpha} = 0.001$$

である。 (σ^2) の推定値は以下の議論に直接の関係がないので省いた)

α の推定値が0.001である。(尤度関数の最大化が非線形最適化の問題になるので、 α に関する格子探索法を用いた。だから“きりの良い”値になっている。)この値は0.8とか0.7くらいになると

思っていたから0.001になったのを見て驚いた。

$\alpha = 0.7$ くらいと予想したのは次の理由による。モデル(2)は“出席者数が減少傾向にある”ことの表現のつもりである。初日の出席者数($\mu_0 + \mu_1 \alpha$)人のうち、 μ_0 人は意志強固な人たち、($\mu_1 \alpha$)人がいわば“浮動票”で、講義がつまらなければこなくなる人、と思っていた。 $\alpha = 0.7$ はほぼ2日で“浮動票”が半分になることを意味する。講義がつまらなくて出てこなくなるパターンというのはこんなものだと予想していたのである。

$\hat{\alpha} = 0.001$ というのは、出席者数の変動のパターンが上記の予想とはまったく異なっていることを示している。

(8)式の値によれば、“浮動票” $\mu_1 \alpha^i$ の値は $i = 1, 2, \dots$ に対して $\{4.2, 0.0, 0.0, \dots\}$ となる。このパターンはどう見ても“講義の様子を見て、つまらなければ来るのをやめる”というパターンではない。これらの人たちも決して安くない授業料を前もって払い込んできているのである。それにもかかわらず1日目の講義だけを聞いてあとをきっぱりとやめている。不思議な行動と言わねばならない。

今“これらの人たち”という言い方をした。まるで初日の講義に出席した78名の人たちのうち、4.2人が“これらの人たち”であり、残りの73.8人が“その他の人たち”であるような言い方である。この言い方は許されるだろうか。毎日ちがう人が何らかの事情で休んで、たまたま表1のような結果になったと考えたほうが良いのではないか。

しかし、もし、毎日ちがう人がそれぞれの事情で休んだのだったとしたら、モデル(1)のAICがモデル(2)のAICより小さく出てしかるべきである。AICの比較によれば、初日の人数が多いのは“偶然”ではないと思われる。

では初日に特に人数の多い“当然”の理由があり得るか。初日にのみあって後の4日にないことはなかったか。講義の内容は何だったか。

表 2 出欠簿 (8日目まで)

受講生番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41									
8/22	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○						
23	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○					
24	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○				
25	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○			
26	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○		
27	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	
29	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	
30	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	計*									
8/22	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○		
23	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
24	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
25	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
26	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
27	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
29	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
30	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

* 表1の集計とちがう。集計に誤りがあったらしい。

1日目の最初の講義を受けもったのは情報量規
準AICの創始者である赤池氏だった。そしてわ
れわれの講座全体はAICの理論と応用に関する
ものだった。これが“当然”の理由かもしれない。

2日目以降の講義に赤池氏は顔を出していない。

もし、以上の想像があたっているとすれば、初
日にのみ出席して、あとはまったく出てこない人
がいるはずである。たぶん4人ほどいると思われ
る。

表2に8日目までの出欠簿を示す。○が出席の
マークである。受講生番号59, 78, 82の所を見て
いただきたい。4人ではなかったが、初日に出席
して2日目以降ずっと顔を出さない人がたしかに
3人いる。“これらの人たち”には実体があったの
である。そしてどうやら“これらの人たち”とい
うのは“赤池ファン”であるらしい。

59番の人は8日目にもう一度顔を見せている。
われわれは講義をしながら、質問を紙に書いて提
出してもらうことにしていた。8日目の午後をそ
の質問に対して解答する時間にあてておいたので

ある。この時には赤池氏も出席した。その日に59
番の人がきている。偶然の一致だろうか。

表2は8日目までしかない。このデータを手に
入れたのが講座の9日目の朝だったからである。
今、この稿に書いているこの話を、9日目の午後
の講義で、AICを用いたデータ解析の1つの例
として話したのである。好評だったように思う。

4. ま と め

以上の話のすじをもう一度整理すると：

- a) 出席者は減少しつつあるのか知りたかつた。
- b) 疑問をモデル選択問題として定式化した。
- c) AIC最小化によって「モデル選択」したところ“減少しつつある”モデルが選ばれた。
- d) パラメータの値をよく見ると、出席者は次第に減っているのではなく、2日目以降が1様に4人ほど少ないという結果である。
- e) 状況をよく考えてみると、出席者の中に特殊な人が4人いて、これらの人が2日目以降

は欠席したということが大いにありそうだ。

f) そこで出欠簿を調べてみると、初日のみ出席して、あとは欠席の人が3人いた。

ということになる。

a) と b) が“モデル選択の前”で、d), e), f) が“モデル選択の後”である。

はっきりいって表1のデータはたいして意味のあるデータではない。われわれの講座に初日だけ出席した人がいようとまいと天下国家の情勢には無関係である。この例に意味があるとすれば、それはこの例が“モデル選択の後”の重要性を示しているからである。

“モデル選択の前”の作業は日常語から統計学語への翻訳であると最初に述べた。“モデル選択の後”の作業は統計学語から日常語への翻訳といってよいだろう。

われわれの例ではこの逆翻訳の過程で妙なことがおこったのである。

“モデル選択の前”には、モデル(2)は“出席者が減りつつある”ことの統計学語訳のはずだった。ところが、モデル(2)が「選択」されてこれを逆翻訳してみたら、“初日の出席者数が異常に多い”ということになってしまった。

要するに α の推定値が予想外の値をとったということである。 $\hat{\alpha}$ が0.7程度でなく0.001になったことを文学的に“データが語った”といっけないうらやまがある。少なくとも筆者がこの値を見た時にそう感じたのは本当である。たった5つの数字が立派に自己主張をしていると思った。

データの解析を始める前には、何らかの予想もっているのが普通である。解析してみても、やはり予想どおりの結果が出たときにはもちろんうれしい。しかし、予想しなかった、そして何か意味のありそうな結果が出てきた時のうれしさというのはまた格別なものである。

この“格別なうれしさ”というのは、つまりは発見の喜びである。発見の喜びは発見された事実が何であるかに無関係ではありえないが、発見さ

れた事実そのものより、発見の仕方に大きく依存するように思われる。

たとえば、筆者が最初に手にしたデータが表1の集計でなく、表2の出欠簿だったとする。そんなことはありそうにないが、この出欠簿をくまなく調べたとしたら、普通と異なる行動様式をとっている3人の受講生に気がついたにちがいない。表1から出発したのと同じ事実は発見できることになる。しかし、この場合には、この発見を面白いと思って長々と原稿に書く気にはならなかったと思う。

“良い”論文というのは、たぶん、1つの明白な命題と、これを言うために必要かつ十分な根拠と論理だけから成り立っているものをいうのだろう。ある“つもり”ではじめた研究が、データにひきずられて思いもかけぬ結果に終わったという論文にはあまりお目にかからないように思う。

統計学という学問が、最も力を発揮するのは、実はこの発見の過程でのことであるはずである。よくあるのは、発見した結果を正当化するために統計学を利用するタイプである。これでは統計学の有効性と面白さは伝わらない。

たった5つの数字をもとに大げさなことを書かせていただいた。しかし、統計学が本当は面白い学問であるというのは本当である。

参 考 文 献

- [1] Akaike, H.: Information theory and an extension of the maximum likelihood principle, 2nd Inter. Symp. on Information Theory (Petrov, B. N. and Csaki, F. eds.), Akademiai Kiado, Budapest, 1973, pp. 267-281.
- [2] 赤池弘次: 情報量規準 AIC とは何か, 数理科学, No. 153 (1976) pp. 5-11.
- [3] 赤池弘次: モデルによってデータを測る, 数理科学, No. 213 (1981) pp. 7-10.
- [4] 坂元慶行, 石黒真木夫, 北川源四郎: 情報量統計学, 共立出版, 1983.